

Getting Started at the OLCF



Presented by:

Bronson Messer

Scientific Computing Group

**Oak Ridge Leadership Computing Facility (OLCF)
National Center for Computational Sciences (NCCS)**



U.S. DEPARTMENT OF
ENERGY

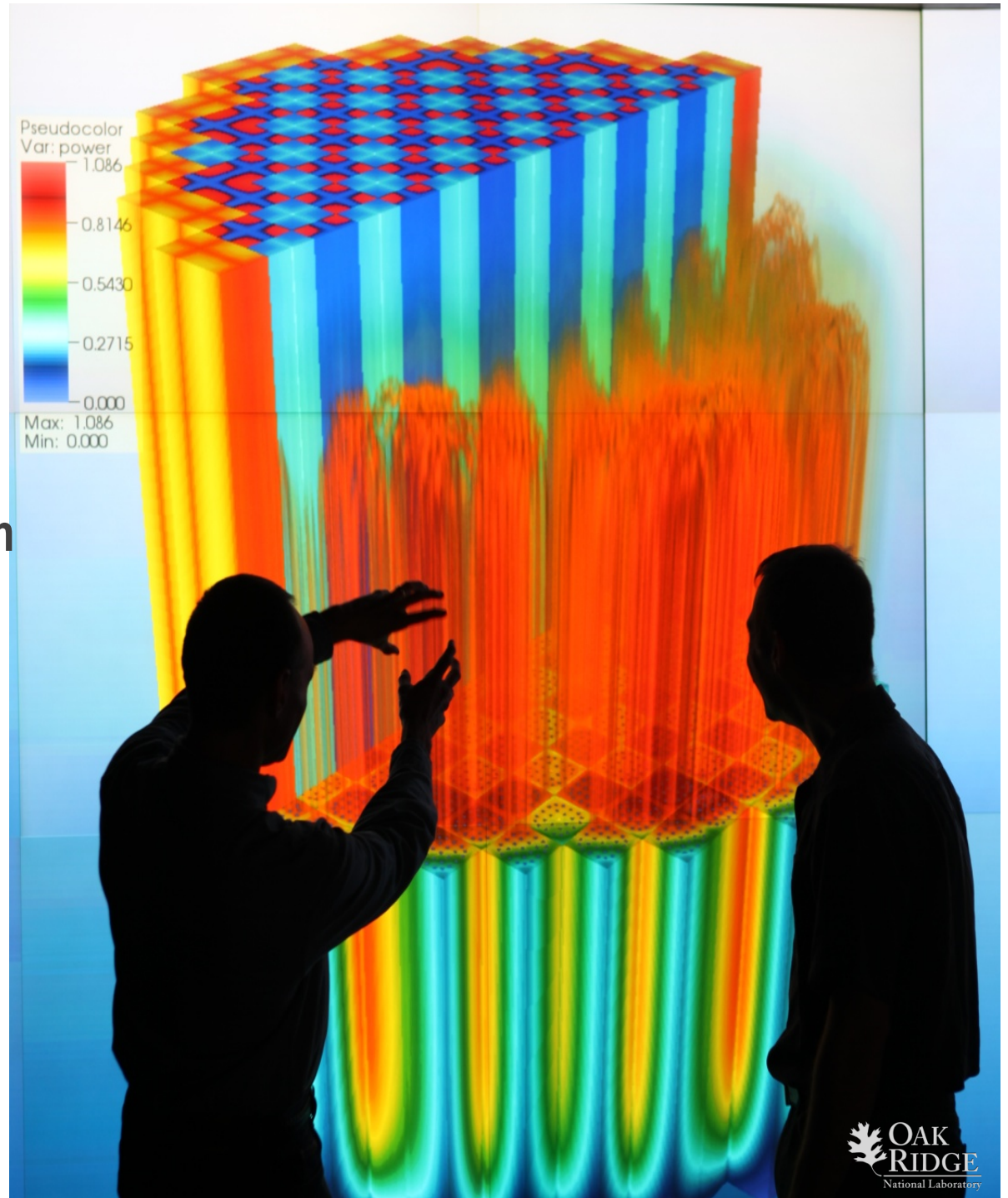


OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

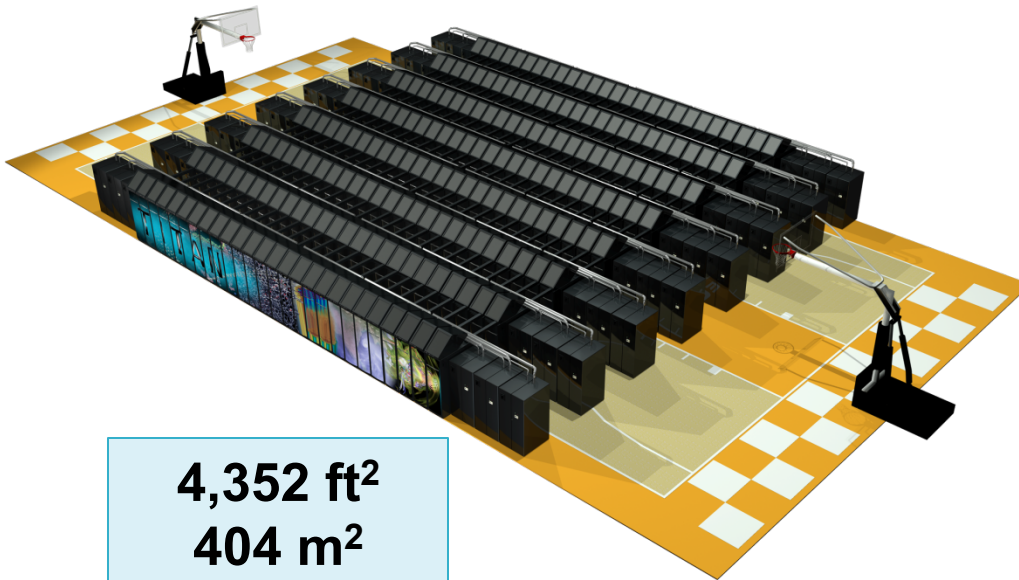
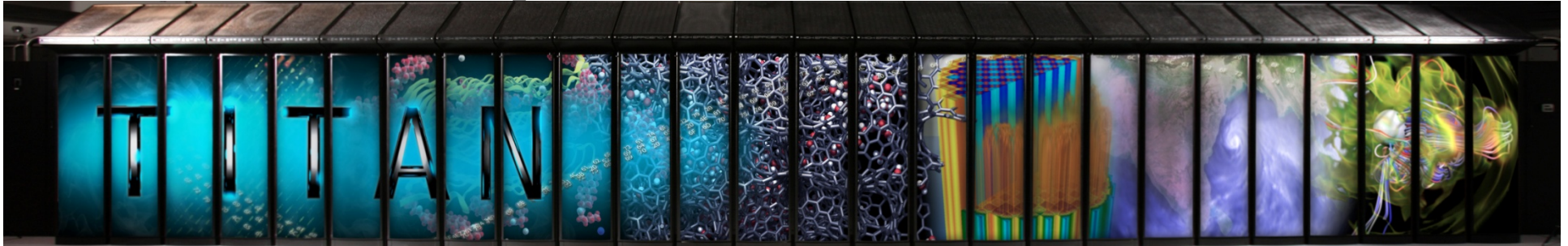
What is Titan?

- The next phase of the Leadership Computing Facility program at ORNL
- An upgrade of Jaguar from 2.3 PF to 27PF
- Built with Cray's newest XK7 compute blades



ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors

#1



4,352 ft²
404 m²

SYSTEM SPECIFICATIONS:

- Peak performance of 27.1 PF
 - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
 - 16-Core AMD Opteron CPU
 - NVIDIA Tesla "K20x" GPU
 - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.8 MW peak power

Cray XK7 Compute Node

XK7 Compute Node Characteristics

AMD Opteron 6200 Interlagos
16 core processor @ 2.2GHz

Tesla M2090 @ 665 GF with
6GB GDDR5 memory

Host Memory
32GB
1600 MHz DDR3

Gemini High Speed Interconnect

Upgraded to NVIDIA's
next generation KEPLER
processor in 2012

Four compute nodes per XK7
blade. 24 blades per rack



Titan “speeds and feeds”

Jaguar specs (2011)

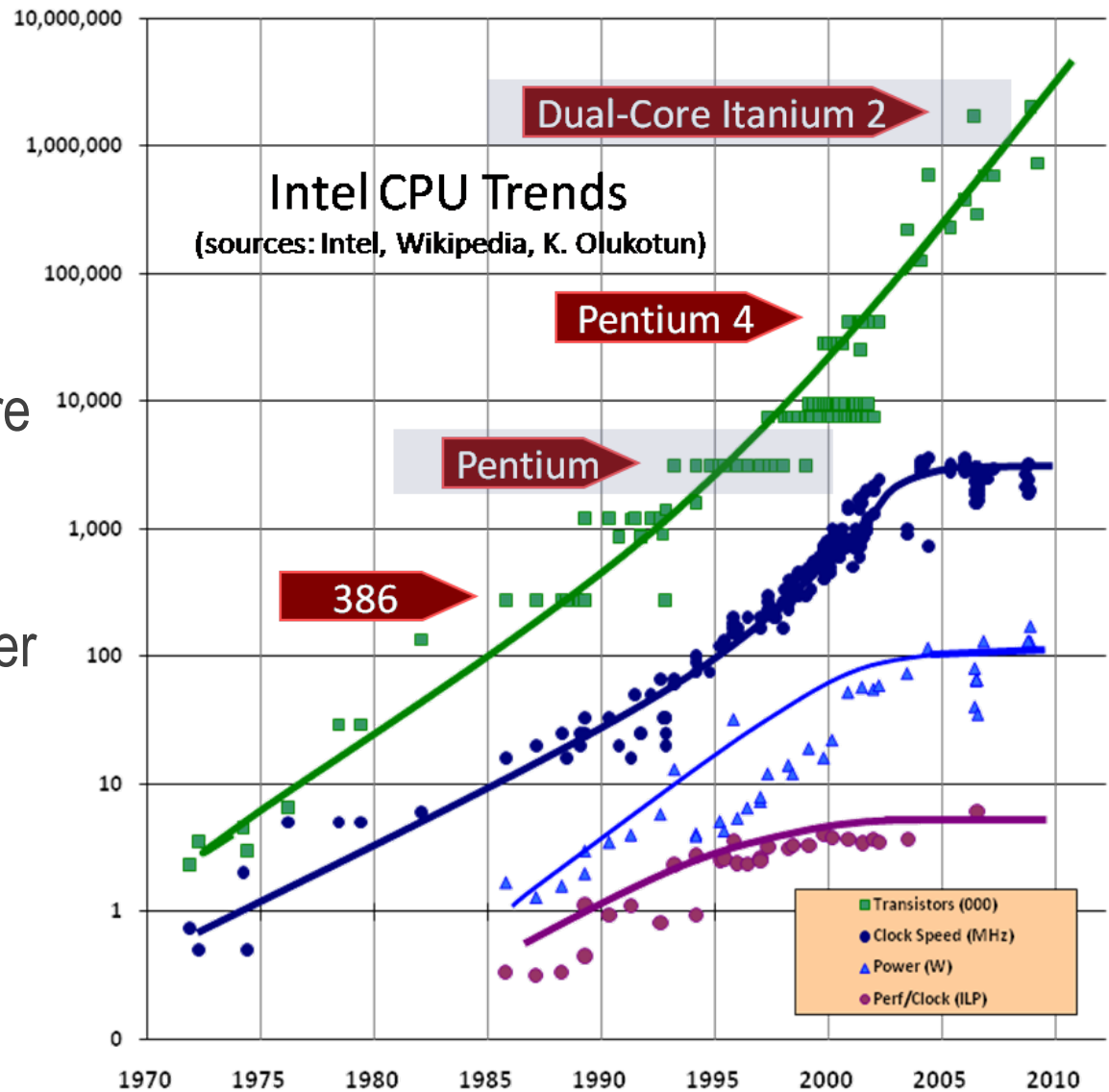
Compute nodes	18,688
Login & I/O nodes	512
Memory per node	24 GB
# of Opteron cores	224,256
# of NVIDIA K20 “Kepler” processors (2013)	NA
Total system memory	450 TB
Total system peak performance	2.3 petaflops

Titan specs (2013)

Compute nodes	18,688
Login & I/O nodes	512
Memory per node	32 GB + 6 GB
# of Opteron cores	299,008
# of NVIDIA K20 “Kepler” processors (2013)	18,688
Total system memory	710 TB
Total system peak performance	27 petaflops

Architectural Trends – No more free lunch

- CPU clock rates quit increasing in 2003
- $P = CV^2f$
Power consumed is proportional to the frequency and to the square of the voltage
- Voltage can't go any lower, so frequency can't go higher without increasing power
- Power is capped by heat dissipation and \$\$\$
- Performance increases have been coming through increased parallelism



Herb Sutter: Dr. Dobb's Journal:

<http://www.gotw.ca/publications/concurrency-ddj.htm>

The Effects of Moore's Law and Slacking¹ on Large Computations

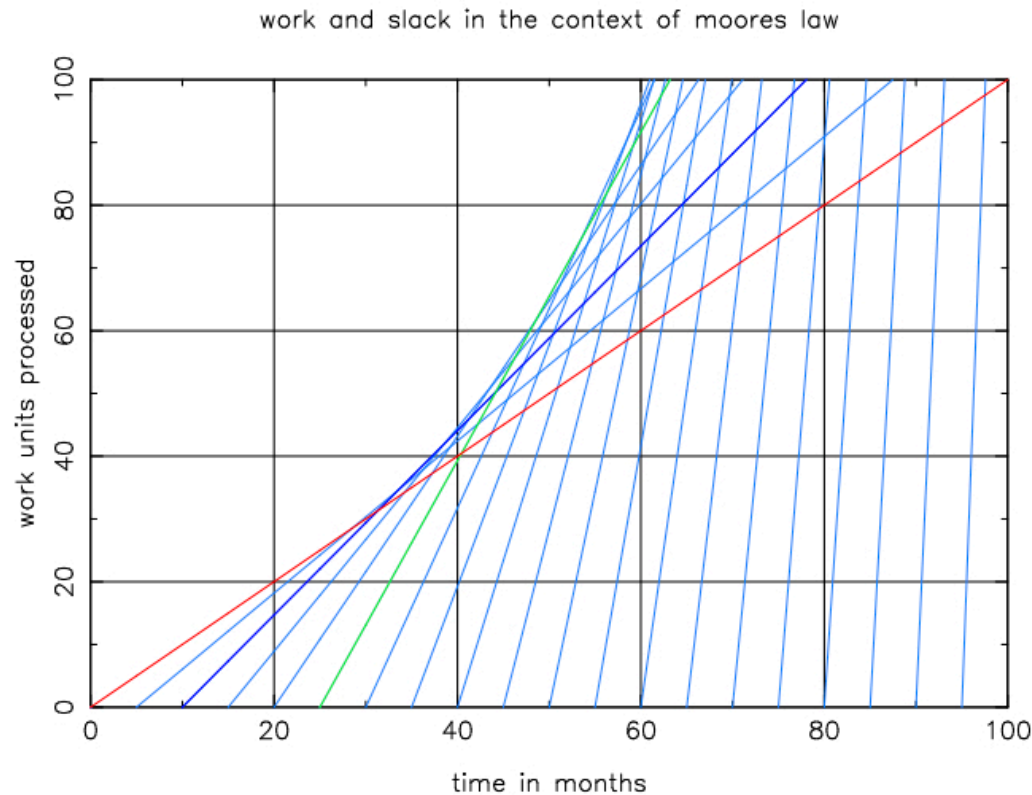
Chris Gottbrath, Jeremy Bailin, Casey Meakin, Todd Thompson,
J.J. Charfman

Steward Observatory, University of Arizona

¹This paper took 2 days to write

Abstract

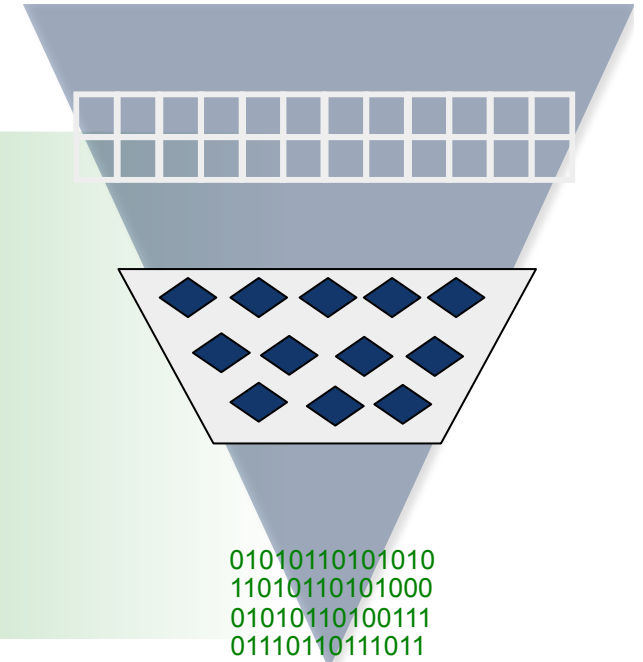
We show that, in the context of Moore's Law, overall productivity can be increased for large enough computations by 'slacking' or waiting for some period of time before purchasing a computer and beginning the calculation.



astro-ph/9912202

Hierarchical Parallelism

- MPI parallelism between nodes (or PGAS)
- On-node, SMP-like parallelism via threads (or subcommunicators, or...)
- Vector parallelism
 - SSE/AVX/etc on CPUs
 - GPU threaded parallelism

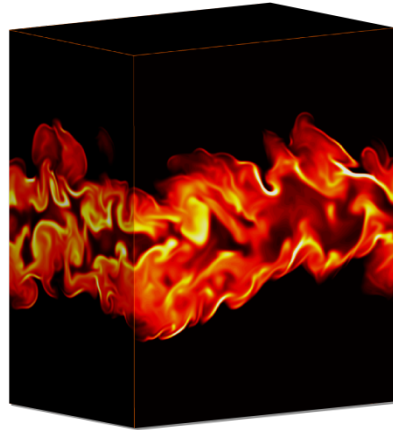


- Exposure of unrealized parallelism is essential to exploit **all** near-future architectures.
- Uncovering unrealized parallelism and improving data locality improves the performance of even CPU-only code.
- Experience with vanguard codes at OLCF suggests 1-2 person-years is required to “port” extant codes to GPU platforms.
- Likely less if begun today, due to better tools/compiler

Titan: Early Applications & Stretch Goals

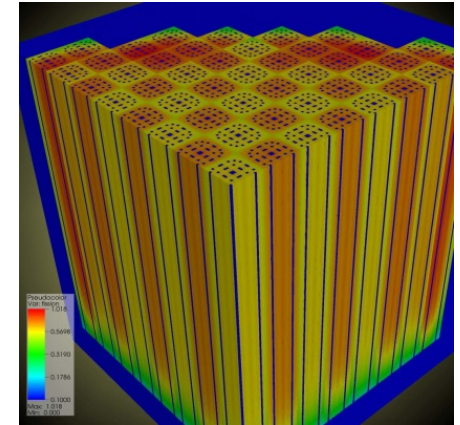
S3D: Turbulent Combustion

Directly solves Navier-Stokes equations. Stretch goals is to move beyond simple fuels to realistic transportation fuels, e.g., iso-octane or biofuels



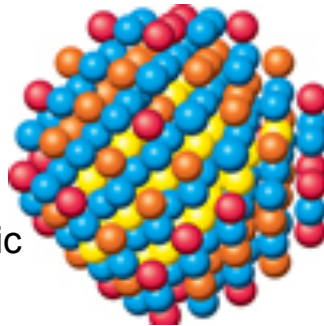
DENOVO: Neutron Transport in Reactor Core

DENOVO is a component of the DOE CASL Hub, necessary to achieve CASL challenge problems



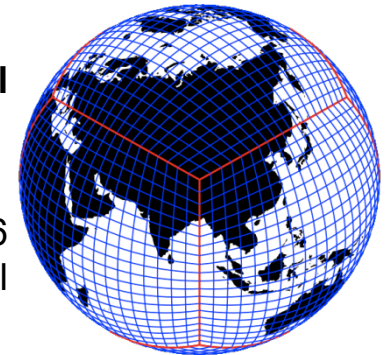
WL-LSMS: Statistical Mechanics of Magnetic Materials

Calculate the free energy for magnet materials. Applications to magnetic recording, magnetic processing of structural materials



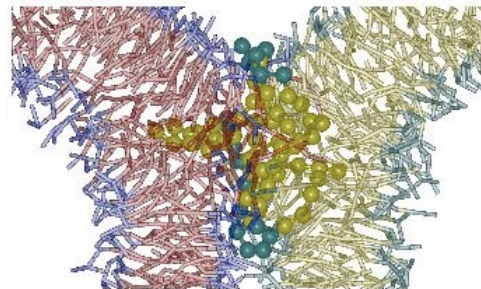
CAM-SE: Community Atmosphere Model – Spectral Elements

CAM simulation using Mozart tropospheric chemistry with 106 constituents at 14 km horizontal grid resolution



LAMMPS: Biological Membrane Fusion

Coarse-grain MD simulation of biological membrane fusion in 5 wall clock days.



How Effective are GPUs on Scalable Applications?

OLCF-3 Early Science Codes -- Performance Measurements on TitanDev

Application	XK6 vs. XE6 Performance Ratio Titan Dev : Monte Rosa
S3D Turbulent combustion	1.4
Denovo 3D neutron transport for nuclear reactors	3.3
LAMMPS Molecular dynamics	3.2
WL-LSMS Statistical mechanics of magnetic materials	1.6
CAM-SE Community atmosphere model	1.5

Cray XK6: Fermi GPU plus AMD 16-core Opteron CPU

XE6: 2X AMD 16-core Opteron CPUs

Additional Applications from Community Efforts

Current performance measurements on TitanDev

Application	XK6 vs. XE6 Performance Ratio Titan Dev : Monte Rosa
NAMD High-performance molecular dynamics	1.4
Chroma High-energy nuclear physics	6.1
QMCPACK Electronic structure of materials	3.0
SPECFEM-3D Seismology	2.5
GTC Plasma physics for fusion-energy	1.6
CP2K Chemical physics	1.5

Running on Titan



U.S. DEPARTMENT OF
ENERGY



OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Leadership Metric and Scheduling Policy

As a DOE Leadership Computing Facility, the OLCF has a mandate to be used for large, *leadership-class* (aka *capability*) jobs.

To that end, the OLCF implements queue policies that enable large jobs to run in a timely fashion.


- Basic queue priority is set by the time a job has been waiting relative to other jobs in the queue.
- However, we use several factors to modify the *apparent* time a job has been waiting. These factors include:
 - The job's processor core request size.
 - The queue to which the job is submitted.
 - The 8-week history of usage for the project associated with the job.
 - The 8-week history of usage for the user associated with the job.

Leadership Usage Metric:

35% of the CPU time used on the system will be accumulated by jobs using 20% or more of the available processors (60,000 cores)

OLCF Scheduling Policy

Bin	Min Nodes	Max Nodes	Max Walltime (Hours)	Aging Boost (Days)
1	11,250	-----	24	15
2	3,750	11,249	24	5
3	313	3,749	12	0
4	125	312	6	0
5	1	124	2	0



Bin 2 is the leadership mark.

NB Nodes are used in the table above; each node “costs” 30 Titan core-hours

OLCF Allocation Overuse Policy

Projects that overrun their allocation are still allowed to run on LCF systems, although at a reduced priority.

- For projects that have used between 100% and 125% of their allocations, the following rules apply:
 - Jobs have their priority reduced by 30 days.
- For projects that have used greater than 125% of their allocation, the following rules apply:
 - Jobs have their priority reduced by 365 days.

To view the entire scheduling policy please see:

http://www.olcf.ornl.gov/kb_articles/scheduling-policy-olcf/

User Support for Titan

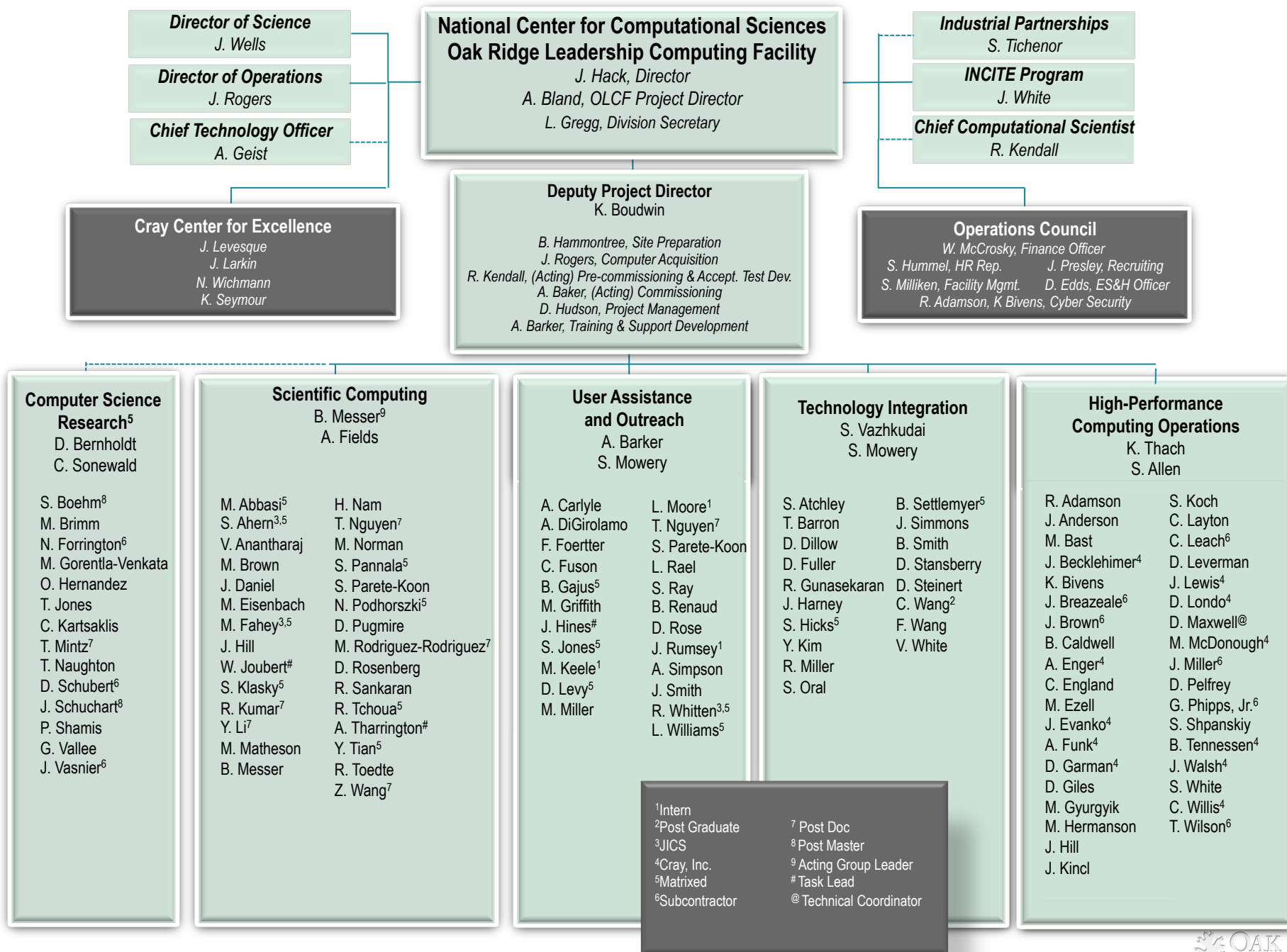


U.S. DEPARTMENT OF
ENERGY



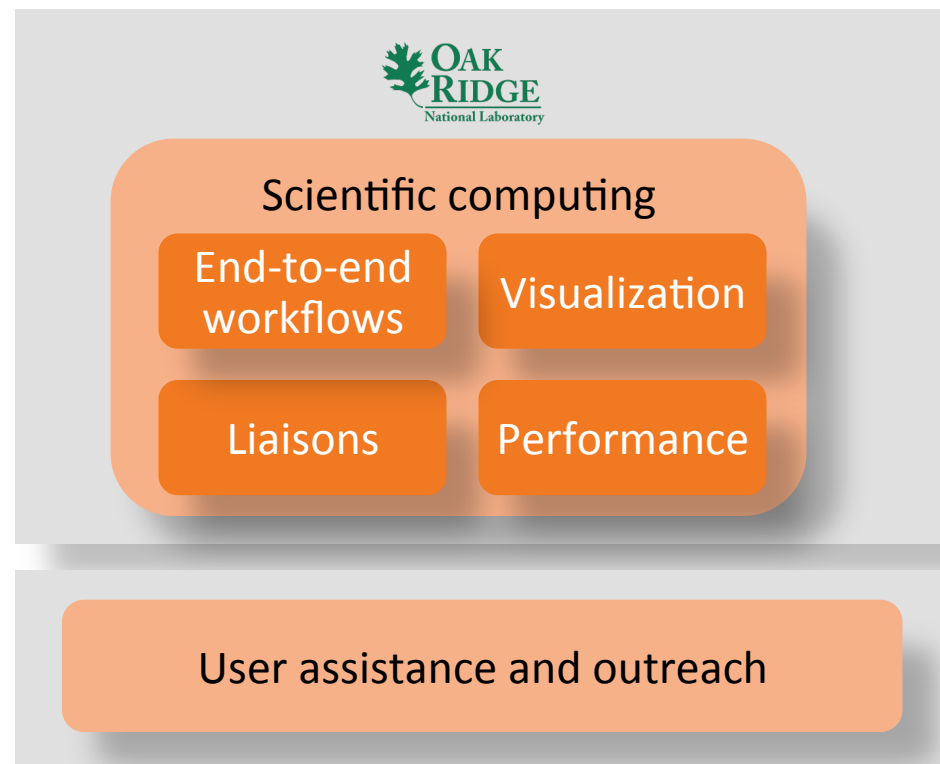
OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY



LCFs support model

- “Two-pronged” support model



Basics

- User Assistance group provides “front-line” support for day-to-day computing issues
- SciComp Liaisons provide advanced algorithmic and implementation assistance
- Assistance in data analytics and workflow management, visualization, and performance engineering are also provided for each project (both tasks are “housed” in SciComp at OLCF)